**Data Statement**

In order to predict genres based on book titles, we compare Naive Bayes and BERT classifiers using a pre-existing dataset from Bhatina (2016) that includes 207,572 books from Amazon, which have been split into 32 different genres. The data we used does not include any personal information or information that could conflict with the GDPR guidelines. The dataset includes information about the authors, genres, titles, and cover images, but for our project, we have included neither authors nor cover images in the data we use. It could be an ethical issue if we tried predicting genres based on author names since predictions might include racial/ethnic biases. With the data we used however, there are neither ethical issues nor dual use problems.

We might be overfitting the data, since we are training the model using “standard” English. If we then test this on titles that incorporate slang or dialect, this might include racial biases (e.g. when testing on titles written in Ebonics). As the data now stands, there is no variety in dialect, since it is written language. We have also chosen to work with only English books, so there is no variety in language either. This means our model would only be applicable to standard written English titles, which don’t include slang or dialect.